ГЛАВА XX. Основные термины и направления исследований

У меня плохая память на всё современное:
ни имен, ни названий, ничего не помню.
А.О. Балабанов

Поговорим о некоторых ключевых терминах, которые имеют отношение к содержанию книги, все они популярны в последние годы, но важно понимать, что за ними скрывается:

- наука о данных (Data Science),
- статистика (Statistics),
- искусственный интеллект (Artificial Intelligence),
- анализ данных (Data Mining),
- машинное обучение (Machine learning),
- большие данные (Big Data).

Наука о данных (Data Science) – это направление науки и технологий представления, сбора, обработки, хранения, анализа и использования данных в цифровой форме, в том числе принятия решения на основе анализа данных. Таким образом, сюда попадают научные направления и технологические решения, которые имеют хоть какое-нибудь отношение к данным, поскольку в определении перечислены все этапы взаимодействия с данными. Все остальные пункты перечня понятий являются названиями разделов «науки о данных», но таких разделов, конечно же больше (теория баз данных, дата-журналистика и т.п.)

Термин «Data Science» появился ещё в 1960х годах, но стал обозначать отдельную дисциплину лишь в начале XXI века.

Анализ данных – нахождение нетривиальных, скрытых, ранее неизвестных и потенциально полезных закономерностей в данных[1]. Важными свойствами этих закономерностей являются:

- **валидность** (они присутствуют в исследуемых данных и отображают некоторые свойства природы данных, а также должны присутствовать в данных, собранных по тем же методологиям, что и исследуемая выборка),
- **потенциальная полезность** (с их помощью обеспечивается экономия времени, ресурсов или возможность заработать деньги),
- **нетривиальность** (закономерности неочевидны до анализа),
- **понятность и интерпретируемость** (описываются в некоторых принятых терминах, и могут быть объяснены специалистам).

Посмотрите фильм «Человек, который изменил всё (MoneyBall)», снятый по реальной истории, описанной в книге: как анализ данных изменил рекрутинговую политику в бейсболе.

В узком смысле термин «анализ данных» определяет сам себя, т.е. перед нами есть данные в понятном нам формате и мы осуществляет анализ – ищем те самые закономерности или преобразуем данные с целью облегчения их поиска или построения модели (данных). В широком смысле – это область человеческой деятельности. Заметим, что «анализ данных» строго говоря не является наукой, это больше ремесло (т.к. необходимо знание и регулярное использования инструментария по обработке данных) и даже искусство (поскольку обнаружение чего-то нетривиального в массиве чисел требует определённой наблюдательности и склада ума). С появлением крупных международных соревновательных платформ типа kaggle.com он превратился даже в спорт, поскольку проводятся соревнования по анализу данных.

Математическая статистика – математическая дисциплина, разрабатывающая математические методы систематизации и использования статистических данных для научных и практических выводов. С ней читатель, наверняка, хорошо знаком, т.к. она входит в обязательные университетские курсы.

Отметим, что теория вероятностей и математическая статистика возникли как прикладные дисциплины, для понимания «случайности» и анализа стратегий в

[1] Термин введён Григорием Пятецким-Шапиро, часто говорят также «интеллектуальный анализ данных», но мы считаем такую конкретизацию лишней, поскольку нет способа оценить «интеллектуальность» анализа и это не соответствует термину на английском языке.

играх. Так первую книгу по этим дисциплинам написал Джероламо Кардано, который систематизировал свой опыт успешного игрока в азартные игры, также на этапе становления большой вклад внёс Блез Паскаль, который также интересовался прогнозированием исходов в играх.

Теперь поговорим о том, что такое **машинное обучение (Machine Learning)**. Для начала заметим, что **обучение** – это приобретение необходимой функциональности посредством опыта, оно бывает двух видов:

- **обучение «на примерах»**, когда опыт приходит некоторыми порциями. Например, так мы учимся ходить: пытаемся сделать шаг, сначала не получается и мы падаем, но по мере накопления опыта мы становимся устойчивее и нам удаётся перемещаться – это и есть нужная нам функциональность. Ещё пример – учим названия животных, в детстве нам их показывают и называют, через некоторое время мы сами безошибочно показываем и называем животных. Здесь нужная функциональность – безошибочно определять название животного по его виду.

- **обучение «по определениям»**, когда одно понятие (или целое знание) определяется через другие (или описывается на специальном языке). Так нас учат в школе на уроке математики: нам дают определение треугольника, а не показывают примеры (хотя без примеров усвоение существенно усложняется), поскольку важно передать точно знание «что такое треугольник». Такое обучение нас в дальнейшем не будет особо интересовать, в науке подобное обучение рассматривается, например, в «**model based reasoning**» – моделировании с помощью уравнений, которые формализуют наше знание о мире (например, в виде дифференциальных уравнений, описывающих физические законы). Мы будем оставаться в рамках «**case based reasoning**» – моделировании на основе прецедентной информации (которая также будет называться «выборкой»).

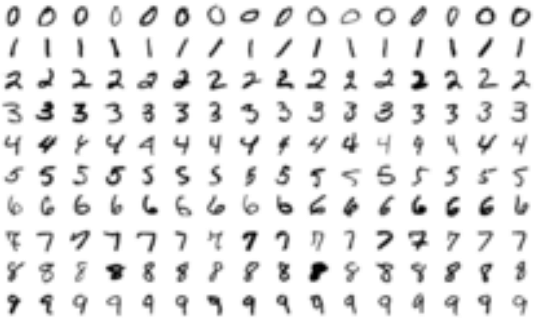
Машинное обучение — процесс, в результате которого машина способна показывать поведение, которое в нее не было явно запрограммировано[1], а также раздел науки, который изучает создание таких процессов. Заметим, что это

[1] Определение из A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229

именно научный раздел (здесь есть теоремы и доказательства, по машинному обучению написаны учебники и задачники).

Важно отметить, что программирование тоже занимается тем, что создаёт процессы с нужной функциональностью, но здесь эта функциональность прописывается явно. Например, при нажатии на клавишу «А» на экране появляется всплывающее окно «Открыть». В машинном обучении же мы программируем систему, которая сама приобретает функциональность (примеры будут дальше).

Ещё одно **определение машинного обучения** – «Компьютерная программа обучается из опыта E в классе задач T с мерой качества P, если качество измеренное с помощью P в классе задач T увеличивается по мере увеличения опыта E[1]». Здесь можно привести несколько примеров, см. ниже



Задача распознавание символов: на вход алгоритму подаётся изображение символа, на выходе символ, соответствующий поданному изображению.

Мера качества: доля правильно распознанных символов (верных ответов).

Опыт: набор изображений символов, размеченных вручную.

Задача обучения логической игре: научить алгоритм играть в логическую игру (шашки / шахматы / го): делать корректные ходы ведущие к победе.

Мера качества: процент побед в играх с фиксированным множеством соперников.

[1] Определение Тома Митчела.

Опыт: каталог игр или игры программы против себя[1].

Задача рекомендации товаров: показывать пользователю на странице товаров ленту рекомендаций (перечень рекомендуемых товаров).

Мера качества: процент успешных рекомендаций (на которые пользователь как-то отреагировал, например посмотрел, положил в корзину или купил).

Опыт: логи действий пользователей (какие товары просматривали, какие рекомендации видели, по каким ссылкам переходили и т.п.).

Можно привести ещё много примеров задач машинного обучения:

- диагностика болезней, прогнозирование эффективности лекарства,
- распознавание образов, символов (Character/ Handwriting Recognition),
- распознавание речи (Speech Recognition),
- распознавание лиц (Face Detection),
- классификация спама (Spam Filtering),
- идентификация (Person Identification / Authentication) лица, отпечатков, радужка глаза и т.п.,
- тональность текста (Sentimental Analysis),
- прогноз спроса / выручки (Demand Forecasting),
- скоринг (Credit Scoring) – определение кредитоспособности,
- определение суммы / пакета страхования,

[1] Сейчас компьютерные программы выучиваются играть лучше человека играя сами с собой, см. алгоритм AlphaZero от компании DeepMind.

- определение психотипа по профилю соцсети / фотографии,
- предсказание оттока (ухода сотрудника / абонента),
- поиск кандидатов на вакансии,
- рекомендации товаров,
- ранжирование Web-страниц,
- ожидание прибыли магазина / рейтинга фильма / доходности сделки,
- анализ форумов, поиск оскорблений, жалоб, автоматическая модерация,
- предсказание поведения клиента / пользователя (например, трат клиента),
- поиск похожих объектов, документов, событий (например, юридических дел),
- обнаружение нетипичных пользователей, фрода, инсайдеров,
- нахождение зависимостей,
- сегментация изображений,
- тегирование/аннотирование документов (automatic summarization).

Приведём пример задачи классификации, которая является особой задачей машинного обучения. Это широко известная **задача классификации ирисов Фишера**[1].

Рис. XX.1. Три вида ирисов: setosa, virginica, versicolor.

Есть три вида ирисов, см. рис. XX.1, статистика по некоторым представителям видов приведена в табл. XX.1.

[1] Рональд Фишер использовал этот набор данных в своей статье.

Длина чашелистника	Ширина чашелистника	Длина лепестка	Ширина лепестка	Вид ириса
4.3	3.0	1.1	0.1	setosa
4.4	2.9	1.4	0.2	setosa
4.4	3.0	1.3	0.2	setosa
				…
4.9	2.5	4.5	1.7	virginica
5.6	2.8	4.9	2.0	virginica
				…
5.0	2.0	3.5	1.0	versicolor
5.1	2.5	3.3	1.1	versicolor

Табл. XX.1. Матрица данных в задаче с ирисами.

Необходимо по этой таблице разработать алгоритм, который осуществляет классификацию, т.е. определяет принадлежность объекта (ириса, заданного своим описанием) к определённому классу (виду ириса).

Ещё пример задачи классификации (а, следовательно, и машинного обучения) – **банковский скоринг**. Когда люди приходят в банк за кредитом, решение о выдачи кредита рекомендуется компьютерной программой (банковский работник лишь подтверждает его). Она анализирует описания всех клиентов банка, для которых известно, вернули ли они кредит. Такие описания можно представить таблицей аналогично тому, как мы описывали ирисы. Только теперь по строкам записаны клиенты банка, а по столбцам – их признаки.

id	статус	г.р.	пол	офис	на счету	просрочки	возврат
43223	физ	1967	М	54	10000	0	Да
43224	физ	1970	Ж	33	2000	2	Нет
43225	юр	1954	М	54	23500	0	Да

Табл. XX.2. Матрица данных в задаче кредитного скоринга.

Заметим, что по математической постановке задача классификации ирисов и клиентов банка совпадают. В обеих задачах данные задаются в табличной форме, один из столбцов в таблице выделен – его называют «целевым». Необходимо научиться в каждой строке определять значения целевого столбца по значениям остальных столбцов.

Большие данные (Big Data) – технологии сбора, хранения, обработки и анализа данных огромных объёмов и значительного многообразия. Это больше

коммерческий и технологический термин, который появился, когда «данных стало много», их стало технологически возможно хранить и обрабатывать, и компании стали это делать для более эффективного ведения бизнеса.

Принято упоминать, то большие данные определяются «несколькими V»:

Velocity – скорость поступления,

Volume – объёмы,

Variety – разнообразие.

Иногда добавляют Veracity (достоверность), Value (ценность) и т.д. При наличии трёх первых «V» объём данных принципиально не уменьшаем. Например, если данные однообразны, например показания датчиков, то можно для каждого датчика описать показания простой моделью, что существенно уменьшит общий объём данных; а если данные не обновляются быстро, то можно провести анализ «по частям». Если выполнены все три «V», то подобные простые приёмы невозможны.

Причины возникновения такого явления как «большие данные» следующие:

- удешевление средств хранения,
- ускорение средств обработки,
- миниатюризация устройств (смартфоны, датчики и т.п.),
- новые форматы / неструктурированность,
- новые технологии (например, GPS),
- интерес бизнеса (появились примеры успешного применения технологий для увеличения эффективности бизнеса),
- успехи отдельных подходов в машинном обучении (например, успехи глубокого обучения).

В качестве примера технологии больших данных можно привести проект Google Flu Trends[1]. Раньше в США эпидемии определялись по анализу отчётов клиник, поэтому об эпидемиях официально заявляли

Об этом и других проектах читайте в книге Виктора Майер-Шенбергера и Кеннета Кукьера «Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим»

[1] https://en.wikipedia.org/wiki/Google_Flu_Trends

часто на 10 дней позже их возникновения. Компания Google научилась определять эпидемии в режиме реального времени, анализируя поисковые запросы. При заболевании люди часто делают поисковые запросы вида «что делать при высокой температуре», «как вылечить насморк» и т.п. Исследователи из Google взяли историю поисковых запросов, сопоставили её со статистикой эпидемий и нашли нужные корреляции. Кроме того, была ещё разработана система прогнозирования эпидемий, которая несколько лет успешно работала, но потом была закрыта, видимо, из-за проблемы «самоотменяющихся прогнозов» (далее мы об этом поговорим подробнее).

Отметим, что поисковые запросы, с помощью которых можно детектировать эпидемию не определялись экспертно (хотя такой вариант решения тоже возможен). Здесь как раз был применён «автоматический анализ данных», который потенциально может обнаружить и нетривиальные закономерности – в данном случае характерные запросы. Например, когда люди болеют, реже совершают дальние поездки, поэтому уменьшается поиск отелей для отдыха.

Искусственный интеллект (ИИ, Artificial Intelligence) – свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека. Также искусственным интеллектом называют эти интеллектуальные системы (программы) и машины, в которых они реализованы (например, роботов), а самое главное – науку и технологию создания этих интеллектуальных систем.

Нужно только конкретизировать, что мы понимаем под «творческими функциями». В широком смысле это все функции, за которые отвечает наш мозг:

- логическое мышление (понимание противоречий, умение делать выводы),
- креативные (сочинять истории и музыку, рисовать и т.п.),
- разговорные (понимать речь, отвечать на вопросы, поддерживать диалог и т.п.),
- ориентация (планирование маршрута, узнавание знакомых мест и т.п.),
- координация (ходьба, бег, акробатические упражнения)
- и т.п.

Несколько примеров ИИ-продуктов:

- умные чат-боты (поддерживают с вами разговор «как человек»),
- автомобили-беспилотники (позволяют безаварийно довезти вас по нужному маршруту),
- умный дом (системы слежения за безопасностью, климат-контроля, выполнение голосовых команд и т.п.).

В качестве одной из первых ярких иллюстраций ИИ обычно приводят компьютер Watson компании IBM, который выиграл в Jeopardy (американский аналог «Своей игры») – передаче, в которой нужно отвечать на интеллектуальные вопросы. Обычно в такой игре участвуют 3 человека, в данном случае компьютер сражался с двумя знатоками, см. рис. XX.2.

Рис. XX.2. Фрагмент игры Jeopardy.

Стоит отметить, что ИИ различают двух видов: в слабом и сильном смыслах. В **слабом смысле** это просто имитация конкретной творческой деятельности человека. Здесь мы наблюдаем постоянную **проблему «почти реализации».** Например, раньше считалось, что только человек способен рисовать красивые картины, сейчас компьютер рисует такие картины, что их не отличают от нарисованных человеком. Казалось бы, вот пример ИИ, но в общественном сознании начинает формироваться мнение, что рисовать картины не так уж и сложно, пусть теперь роботы смогут, скажем, обыграть человека в футбол. В **сильном смысле** от ИИ требуются способности мыслить и осознавать себя как отдельную личность (в частности, понимать собственные мысли), т.е. кроме интеллектуального поведения компьютер должен обладать искусственным сознанием:

- самоидентифицировать себя (понимать, что такое «я»),
- идентифицировать других и противопоставлять себе (как ребёнок понимает «мама», «папа», «чужой»),

- бороться за ресурсы (т.е. иметь желания обладать чем-то и реализовывать их, подобно тому, как ребёнок тянется к игрушке).

Также сильный ИИ способен самостоятельно ставить перед собой задачи и находить их решения.

Термины: итоги

- Центральные термины этой главы обозначают разные понятия, хотя их часто и путают (скажем, искусственный интеллект не тождественен машинному обучению, хотя современные ИИ-системы базируются на методах машинного обучения).
- Самым широким термином является «наука о данных», т.к. охватывает все направления работы с данными[1].
- В этой книге мы, в основном, будем говорить о связке машинного обучения и анализа данных.

Спасибо за внимание к книге!
Замечания по содержанию, замеченные ошибки
и неточности можно написать в телеграм-чате
https://t.me/Dyakonovsbook

[1] Рекомендуемая литература:
Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим // М.: Манн, Иванов и Фербер. – 2014. – Т. 240.
Том Таулли «Основы искусственного интеллекта: нетехническое введение» БХВ-Петербург. – 2021.
Домингос П. Верховный алгоритм: как машинное обучение изменит наш мир // М.: Манн, Иванов и Фербер, 2016.